Pattern Recognition 44.7 (2011): 1426 – 1434

DOI: http://dx.doi.org/10.1016/j.patcog.2010.12.021

Inference on the Prediction of Ensembles of Infinite Size

Daniel Hernández-Lobato*,[a], Gonzalo Martínez-Muñoz[b], Alberto Suárez[b]

[a] *Machine Learning Group, ICTEAM Institute, Université catholique de Louvain, Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium.*
[b] *Computer Science Department, Escuela Politécnica Superior, Universidad Autónoma de Madrid, C/ Francisco Tomás y Valiente, 11, Madrid 28049 Spain.*

Abstract

In this paper we introduce a framework for making statistical inference on the asymptotic prediction of parallel classification ensembles. The validity of the analysis is fairly general. It only requires that the individual classifiers are generated in independent executions of some randomized learning algorithm, and that the final ensemble prediction is made via majority voting. Given an unlabeled test instance, the predictions of the classifiers in the ensemble are obtained sequentially. As the individual predictions become known, Bayes' theorem is used to update an estimate of the probability that the class predicted by the current ensemble coincides with the classification of the corresponding ensemble of infinite size. Using this estimate, the voting process can be halted when the confidence on the asymptotic prediction is sufficiently high. An empirical investigation in several benchmark classification problems shows that most of the test instances require querying only a small number of classifiers to converge to the infinite ensemble prediction with a high degree of confidence. For these instances, the difference between the generalization error of the finite ensemble and the infinite ensemble limit is very small, often negligible.

Key words: Classification Ensembles, Classification Trees, Bayesian Inference, Infinite Ensembles

*Corresponding author. Tel: +32-10-47-2445; fax: +32-10-45-0345.
Email addresses: daniel.hernandez-lobato@uclouvain.es (Daniel Hernández-Lobato), gonzalo.martinez@uam.es (Gonzalo Martínez-Muñoz), alberto.suarez@uam.es (Alberto Suárez)

1. Introduction

Ensembles are among the most successful methods used to address supervised learning problems [1, 2, 3, 4, 5, 6, 7]. The prediction of an ensemble is obtained by combining the individual predictions of a collection of diverse classifiers. Provided that these predictions are complementary, ensembles provide an effective mechanism to achieve better generalization performance. In this work we consider parallel ensembles of classifiers of the same type. The individual classifiers in the ensemble are generated in independent executions of a randomized learning algorithm. This procedure takes advantage of instabilities in the base learning algorithm to generate a collection of diverse classifiers [3, 6]. Finally, the prediction of the ensemble is computed by majority voting. Bagging [1], random forest [2], extra-trees [7], subagging [4], rotation forest [6] and class-switching ensembles [5] are representative ensembles of this kind.

In these types of ensembles the generalization error typically decreases as the size of the ensemble increases [1, 2, 8, 7, 5]. In general, the larger the ensemble is, the more accurate its prediction. However, the rate of improvement in performance becomes smaller as the size of the ensemble increases. Furthermore, the computational costs of generation, storage and prediction increase linearly with the number of classifiers included in the ensemble. Therefore, it is important to determine whether it is possible to estimate the prediction of an ensemble of very large size (ideally of infinite size) using only the predictions of a finite collection of classifiers. Or, alternatively, to quantify how confident one can be that the prediction of an ensemble of finite size coincides with the prediction of the corresponding ensemble of infinite size. In this work we show that the answer to these questions strongly depends on the particular instance that is being classified. For most instances, the infinite ensemble prediction can be estimated with a very high degree of confidence using the predictions of only a small number of classifiers. By contrast, instances that are close to classification frontiers (usually a small fraction of the instances considered) require querying a very large number of classifiers to converge to the asymptotic (infinite) ensemble prediction.

These questions can be addressed by analyzing the convergence of majority voting in the infinite-ensemble limit. The probabilistic framework described in [9, 10] is particularly suited for this purpose. For a given instance, the asymptotic prediction of the ensemble can be expressed in terms of the set of probabilities that an individual classifier assigns a particular class label

to that instance. The difficulty is that these class probabilities, which depend on the particular instance considered, are initially unknown. Nevertheless, the voting process provides information that can be used to estimate their distribution. Starting from a uniform prior, Bayes' theorem is used to compute a posterior that incorporates the evidence given by the predictions of the individual classifiers as they become known. The posterior distribution describes the uncertainty of the provisional estimates of the class probabilities. This distribution is then used to compute the probability that the class label currently predicted by the finite ensemble (the current majority class) coincides with the class label that an ensemble of infinite size would predict. Provided that a small amount of uncertainty in the final prediction is acceptable, the voting process can be stopped when the probability estimate exceeds some specified threshold, π. This stopping strategy guarantees that the differences between the classification error of the finite ensemble and of the infinite-ensemble are at most $1-\pi$. This is because the differences in error are necessarily smaller than the differences in class predictions. In particular, if the changes in the class labels affect correctly and incorrectly classified instances in approximately equal numbers, the differences in classification error should be much smaller than this upper bound. The validity of this analysis is illustrated in extensive experiments in benchmark classification problems. In these problems, most of the test instances require knowing the output of only a few classifiers to produce a reliable estimate of the asymptotic ensemble prediction. Furthermore, the error of the ensemble in this subset of test instances is very close to the asymptotic (infinite ensemble) limit.

The organization of the manuscript is as follows: In Section 2 we analyze the prediction process of the ensemble by majority voting. This analysis is used to make inference about the prediction of the ensemble in the infinite-size limit. Section 3 discusses the relation of the present work with analyses found in the literature. In Section 4 the results of experiments in a wide range of classification problems are used to illustrate the validity of the proposed framework. Finally, the results and conclusions of this investigation are summarized in Section 5.

2. Inference on the Asymptotic Ensemble Prediction

Consider an ensemble $\{h_i(\mathbf{x})\}_{i=1}^{t}$ composed of t classifiers. Assuming that majority voting is used to combine the decisions of the individual predictors, the class label assigned by the ensemble to an unlabeled instance described

by the vector of attributes $\mathbf{x}$ is

$$\hat{y}^t = \arg\max_y \sum_{i=1}^{t} \mathcal{I}(h_i(\mathbf{x}) = y),\ y \in \mathcal{Y}\,, \tag{1}$$

where $\mathcal{I}$ is an indicator function and $\mathcal{Y} = \{y_1, \ldots, y_l\}$ is the set of possible class labels.

As described in [10], if the individual classifiers of the ensemble are built independently when conditioned to the training data[1], the polling process defined in Eq.˜(1) can be seen as a sequence of t independent trials. Each individual trial corresponds to the classification of $\boldsymbol{x}$ given by an individual classifier. The possible outcomes of the t independent trials follow a multinomial distribution

$$\mathcal{P}(\mathbf{t}|t, \boldsymbol{p}(\mathbf{x})) = \frac{t!}{t_1!, \ldots t_l!} p_1(\mathbf{x})^{t_1} \ldots p_l(\mathbf{x})^{t_l}\,, \tag{2}$$

where $\mathbf{t} = \{t_1, t_2, \ldots, t_l; \sum_{i=1}^{l} t_i = t\}$, t_i is the number of classifiers that predict class label y_i and $\boldsymbol{p}(\mathbf{x})$ is the probability vector

$$\boldsymbol{p}(\mathbf{x}) = \{p_1(\mathbf{x}), p_2(\mathbf{x}), \ldots, p_l(\mathbf{x})\}, \quad \sum_{i=1}^{l} p_i(\mathbf{x}) = 1. \tag{3}$$

The quantity $p_i(x)$ is the probability that an ensemble classifier assigns the label $y_i \in \mathcal{Y}$ to instance $\mathbf{x}$. The values of these probabilities depend on the ensemble learning algorithm, on the particular classification problem and on the specific instance considered. To simplify the notation, the dependence of $\boldsymbol{p}$ on $\mathbf{x}$ will be assumed implicit in the rest of the article.

If the value of $\boldsymbol{p}$ for instance $\mathbf{x}$ is known, Eq.˜(2) can be directly used to compute the probability that an ensemble of size t assigns the class label y_i to $\mathbf{x}$. One simply needs to sum (2) over all possible situations in which the number of ensemble classifiers that predict the class label y_i is larger than the number of classifiers that predicted any other class label. For binary

[1]Note that this is different from assuming that the classifiers are unconditionally independent.

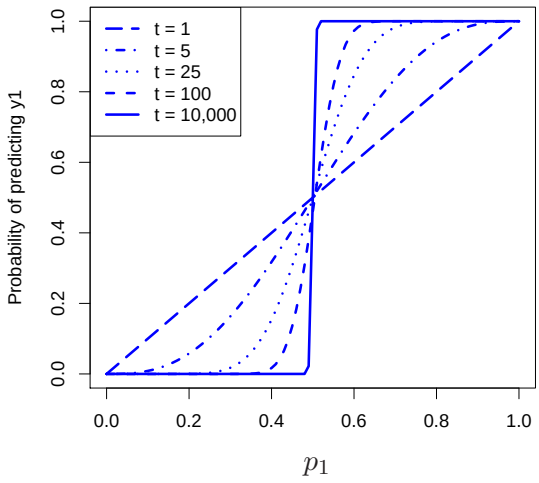

Fig. 1: Probability of predicting class label y_1 as a function of p_1 for different values of t, the ensemble size. In the limit $t \to \infty$ this probability is a step function that takes value 0 when $p_1 < 1/2$ and 1 when $p_1 > 1/2$.

classification problems ($l = 2$) this probability is

$$\begin{aligned} \mathcal{P}(\hat{y}^t = y_1 | t, p_1) &= \sum_{t_1 = \lceil \frac{t}{2} \rceil}^{t} \binom{t}{t_1} p_1^{t_1} (1 - p_1)^{t - t_1} \\ &= I_{p_1}\left(\lfloor \frac{t}{2} \rfloor + 1, t - \lfloor \frac{t}{2} \rfloor \right) , \end{aligned} \tag{4}$$

where $I_p(a, b)$ is the regularized incomplete beta function [11]. For multi-class problems this can be a costly computation because the number of terms that need to be included in the sum grows exponentially with the number of possible class labels l.

In the limit of an ensemble of infinite size ($t \to \infty$), the class prediction probabilities given by Eq.˜(4) become Boolean values. Specifically, the asymptotic class prediction is y_i with $i = \arg\max_k p_k$. Fig. 1 displays the dependence of (4) on p_1, the probability of predicting class label y_1 in a binary classification problem, for different values of the ensemble size t. For $t = 1$, (4) is the identity function. As t grows, it approaches a step function. In the limit $t \to \infty$, the probability that the ensemble predicts class y_1 for $p_1 > 1/2$ tends to one while for $p_1 < 1/2$ this probability approaches zero.

2.1. Inference on the prediction of an ensemble of infinite size

According to the probabilistic framework introduced, the asymptotic prediction of the ensemble for a given instance $\mathbf{x}$ can be computed in terms of the vector of class probabilities $\boldsymbol{p}$. These probabilities are initially unknown. Nevertheless, it is possible to make inference about $\boldsymbol{p}$ using the evidence given by the predictions of a finite number of ensemble classifiers. Suppose that the votes of t classifiers $\boldsymbol{t} = \{t_1, t_2, \ldots, t_l; \quad \sum_{i=1}^{l} t_i = t\}$ are known. Assume a uniform prior distribution for $\boldsymbol{p}$. The multinomial likelihood described in Eq.˜(2) can be combined with this prior to compute a posterior distribution for $\boldsymbol{p}$ using Bayes' theorem

$$\mathcal{P}(\boldsymbol{p}|\mathbf{t}) = \frac{\mathcal{P}(\mathbf{t}|\boldsymbol{p})\mathcal{P}(\boldsymbol{p})}{\mathcal{P}(\mathbf{t})} = \frac{\Gamma(\sum_{i=1}^{l} t_i + l)}{\prod_{i=1}^{l} \Gamma(t_i + 1)} \, p_1^{t_1} p_2^{t_2} \ldots p_l^{t_l} \,, \tag{5}$$

where $\Gamma(z)$ is the gamma function. The posterior is a Dirichlet distribution of order l with parameters $(t_1 + 1, \ldots, t_l + 1)$.

Eq.˜(5) can be used to make inference on the asymptotic prediction of the ensemble when t classifiers have been queried. Specifically, one needs to compute the probability that one component of the vector $\boldsymbol{p}$ is higher than the other components.

$$\mathcal{P}(\hat{y}^{\infty} = y_i|\mathbf{t}) = \mathcal{P}\left(\bigcap_{j \neq i} p_i > p_j|\mathbf{t}\right) . \tag{6}$$

In binary classification problems this probability is

$$\mathcal{P}(\hat{y}^{\infty} = y_1|\mathbf{t}) = \mathcal{P}\left(p_1 > p_2|\mathbf{t}\right) = I_{1/2}(t_2 + 1, t_1 + 1) \,. \tag{7}$$

In multi-class problems, Eq.˜(6) is difficult to compute. However, it can be approximated by a lower bound $\mathcal{L}(y_i|\mathbf{t}) \leq \mathcal{P}(\hat{y}^{\infty} = y_i|\mathbf{t})$ that provides a conservative estimate of the confidence level on the asymptotic prediction

$$\mathcal{P}(\hat{y}^{\infty} = y_i|\mathbf{t}) \geq \mathcal{L}(y_i|\mathbf{t}) = \prod_{j \neq i} \mathcal{P}\left(p_i > p_j|\mathbf{t}\right) = \prod_{j \neq i} I_{1/2}(t_j + 1, t_i + 1). \tag{8}$$

The derivation of this lower bound uses the relation

$$\mathcal{P}(p_i > p_j| \bigcap_{k \in K} p_i > p_k, \mathbf{t}) \geq \mathcal{P}(p_i > p_j|\mathbf{t}) \tag{9}$$

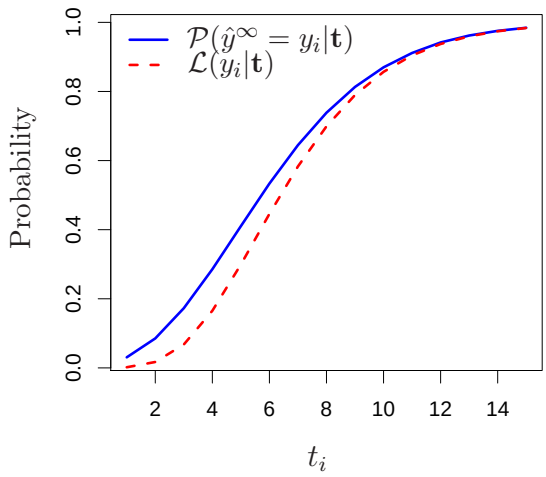

Fig. 2: Plots of the probability $\mathcal{P}(\hat{y}^{\infty} = y_i|\mathbf{t})$ (6) and of the lower bound $\mathcal{L}(y_i|\mathbf{t})$ (8) as a function of t_i, the number of votes for class y_i, $l = 5$ and fixed values of $\{t_j;\ j \neq i\} = \{5, 3, 2, 1\}$.

$\forall K$ in $\{1, 2, \ldots, l\} \setminus \{i, j\}$, which can be derived from the FKG inequality [12]. In binary classification problems the bound $\mathcal{L}(y_i|\mathbf{t})$ coincides with the exact value, given by Eq.˜(7). The proposed lower bound is fairly tight, specially for values of $\mathcal{P}(\hat{y}^{\infty} = y_i|\mathbf{t})$ close to 1. This is illustrated in Fig. 2 for a problem with five classes ($l = 5$). The curves in Fig. 2 display the dependence of $\mathcal{P}(\hat{y}^{\infty} = y_i|\mathbf{t})$ and $\mathcal{L}(y_i|\mathbf{t})$ as a function of t_i, the number of observations of class i, for fixed values of the votes for the other classes $\{t_j;\ j \neq i\} = \{5, 3, 2, 1\}$. Similar curves are obtained for different values of l and of $\{t_j;\ j \neq i\}$. The exact values of $\mathcal{P}(\hat{y}^{\infty} = y_i|\mathbf{t})$ used in this graph are calculated by a standard numerical quadrature algorithm. The time-complexity of the exact calculation is exponential in the number of classes and soon becomes unmanageable as the number of classes increases.

2.2. Stopping criterion

Consider a specific instance to be classified. Assume that the predictions of a finite number of ensemble classifiers are known. These predictions are summarized in the vector of vote counts $\mathbf{t}$. The lower bound of the probability estimate $\mathcal{L}(y_i|\mathbf{t})$ can be used to determine when the evidence given by $\mathbf{t}$ is sufficient to provide an estimate of the asymptotic ensemble prediction with a high level of confidence π. Specifically, if the estimate of the probability $\mathcal{L}(y^*|\mathbf{t})$ for the current majority class, y^*, exceeds threshold π,

the polling process can be stopped. This prescription guarantees that the class label assigned by the finite and by the infinite ensembles coincide with a probability greater than or equal to $\approx \pi$. The number of classifiers that need to be queried in a given classification task depends on the value of π. As π approaches 1, this number diverges.

Consider the instances that satisfy the stopping criterion for a given confidence level π. For these instances, the differences between the classification error of the finite ensemble and the infinite ensemble limit must be smaller than $1 - \pi$. In the classification problems investigated these differences are generally much smaller than $1 - \pi$. This means that the changes in class labels affect correctly and incorrectly classified instances in approximately equal numbers.

The overhead of determining whether the querying process should be halted is small, provided that some computations are made beforehand. Specifically, the evaluation of (8) involves the product of $(l-1)$ terms. Each term is an evaluation of the regularized incomplete beta function on integers. The required computations are identical irrespective of the classification problem or the type of randomized parallel ensemble considered. Thus, the values required for these computations can be precalculated, stored in memory and retrieved when needed. The cost of retrieving $(l-1)$ of these tabulated values, multiplying them and comparing the result to the specified confidence level π is fairly small. In binary classification problems it is more convenient to store the values of $t^*(t;\pi)$, i.e. the minimum number of majority class votes needed to guarantee that the prediction of the ensemble of size t coincides with the asymptotic ensemble prediction.

Fig. 3 illustrates the use of the proposed stopping criterion in a binary classification problem. The curves displayed in this figure correspond to the fraction of class y_1 votes, t_1/t, as a function of the total number of votes t for three different instances. Each instance is characterized by a different value of p_1 (the probability that an arbitrary ensemble classifier assigns class y_1 to that instance). The black curve corresponds to the critical value $t^*(t;\pi = 0.99)/t$. As more predictions are known, the number of class y_1 votes follows a binomial process with parameter p_1. This random walk eventually reaches the critical threshold $t^*(t;\pi)/t$. This time of first passage is marked with a circle in the plots. At this point the confidence on the class estimate is above $\approx \pi$ and the querying process can be stopped. In this figure, one can observe that the instance with $p_1 = 0.5$ has not reached the critical level even after querying 200 classifiers. Instances with p_1 in the

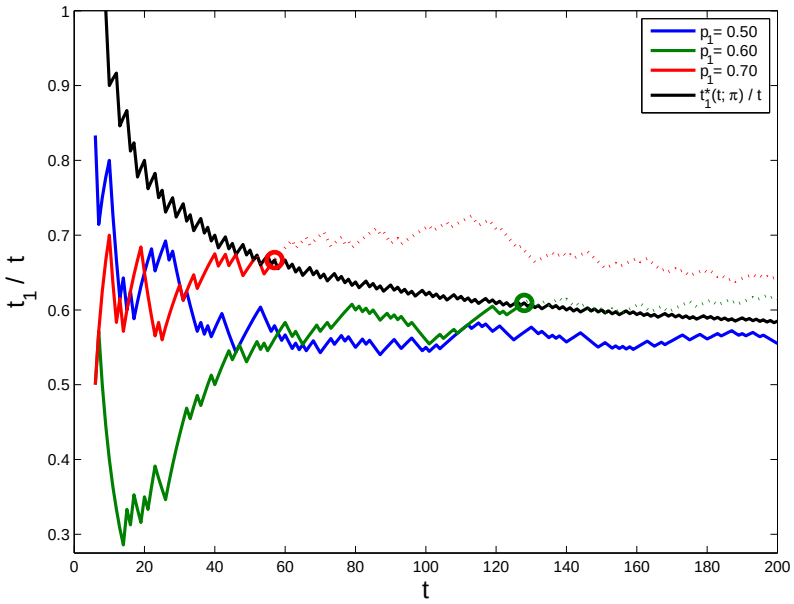

Fig. 3: Prediction by majority voting in a binary classification process for instances with different values of p_1. The fraction of class y_1 votes (t_1/t) is plotted as a function of the total number of votes t. The black line corresponds to the critical value $t^*(t;\pi = 0.99)/t$. When the fraction of majority class votes obtained by an instance are above this line, the probability that the current majority class coincides with the asymptotic majority class is above π.

vicinity of 0.5 may require an extremely large number of classifiers to receive a stable prediction with a high level of confidence.

In typical classification problems only a few instances (those near classification frontiers) have values of p_1 close to 0.5. Most instances reach their asymptotic prediction after querying only a few classifiers. This observation is illustrated in a simple two-dimensional binary classification task. In this problem, instances are drawn from a uniform distribution in the region $[-1, 1] \times [-1, 1]$. Instances inside a circumference of radius $\sqrt{2}$ centered at the origin are class 1. Class label 2 is assigned to instances outside the circle. To make the problem more realistic some noise is injected: the class of 5% of the instances selected at random is flipped. The experiment is repeated 100 times. For each realization we generate a training set of 300 examples.

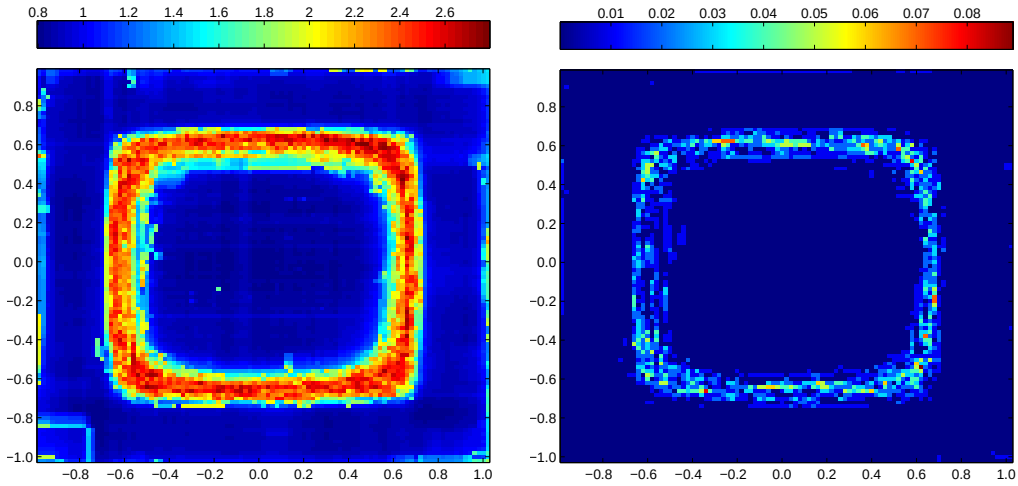

Fig. 4: (*Left*) Average number of classifiers (in $\log_{10}$ scale) used to estimate with $\pi = 0.99$ certainty the class labels given by an ensemble of infinite size. Instances that required more than $10,000$ classifiers to reach a stable classification are not included this plot. (*Right*) Fraction of instances that required more than $10,000$ classifiers to reach a stable classification in the 100 executions.

Each of these training sets is used to generate a bagging ensemble of size $10,000$. The generalization performance of the ensemble is estimated using a test set consisting of 101×101 points located on a regular grid in the region $[-1, 1] \times [-1, 1]$. For each test point, the individual classifiers in the ensemble are queried, one classifier at a time, until the confidence in the class prediction is above $\pi = 0.99$ or until all $10,000$ classifiers have been queried. The results of these experiments are summarized in Fig. 4. On the *left* plot, the average number of queried classifiers is shown as a heat map in logarithmic ($\log_{10}$) scale. To compute this average we have removed instances whose stable classification would have required more than the $10,000$ classifiers generated. The *right* plot shows also as a heat map the fraction of times that more than $10,000$ classifiers would have been needed to reach a confidence level $\pi = 0.99$ in the class assignment. These figures show that instances away from the classification frontier require querying a fairly small number of classifiers. In contrast, the prediction of instances close to the class boundaries between the two classes takes many more classifiers to converge.

3. Related Work

The analysis of the prediction process presented in the previous section applies to any kind of parallel ensemble in which the individual classifiers are generated in independent executions of a randomized learning algorithm. These include bagging [1], random forest [2], extra-trees [7], subagging [4], rotation forest [6] and class-switching ensembles [5], among others. By contrast, the analysis cannot be directly applied to sequential ensemble algorithms, such as boosting [13, 14]. In boosting, the ensemble grows by incorporating new classifiers that focus on instances that are difficult to predict by the classifiers included earlier. This procedure introduces correlations among the classifiers that invalidate the independence assumption.

The analysis of the prediction of an ensemble in terms of majority voting was first made by Hansen and Salamon [9]. In that work, the authors derive explicit formulas for the ensemble error assuming that the prediction errors of the individual classifiers of the ensemble in the test set are independent. These formulas depend on the size of the ensemble and on the probability that an individual classifier of the ensemble predicts the correct label for an arbitrary example. The assumption that the errors of the individual predictors are independent is clearly not realistic. In a more sophisticated analysis they take into account the possibility that the prediction for a specific instance has an associated level of difficulty that is independent of the particular ensemble member that is doing the classification. The infinite ensemble limit is not considered explicitly. Nevertheless, this limit can be readily derived from their formulas.

In [2] the Strong Law of Large Numbers is used to prove that the prediction of random forests converges almost surely to a limit as the number of trees in the forest becomes large. Therefore, the generalization error for random forests converges, which means the infinite ensemble limit is well defined. This proof also applies to randomized ensembles of the type considered in this investigation. This convergence property explains why randomized ensembles do not overfit as more predictors are included.

In [15, 16], ensemble learning and, in particular, bagging and boosting are analyzed from the perspective of Monte Carlo algorithms. The authors introduce a Monte Carlo ensemble learner that extracts hypotheses at random any time it needs to classify a new example. Bagging is shown to be an approximation to such a process. The analysis presented in [16] allows to establish under which conditions the ensemble can be expected to outper-

form the best single predictor. Thanks to the special properties of the Monte Carlo ensemble learner, it is possible to derive a closed form expression for the expected ensemble error as a function of the number of classifiers considered for prediction. The case of an ensemble with an infinite number of classifiers is also explicitly discussed.

The prediction error of majority voting has also been analyzed in terms of classification margins in [17]. The classification margin for a test instance $\mathbf{x}$ is defined as the difference between the fraction of votes given by the ensemble for the correct class and the maximum fraction of votes assigned to some other class label. Assuming that the distribution of these margins is known, [17] derives an upper bound on the ensemble error. This bound provides an estimate of the average ensemble classification error. Using this bound one can also show that, as the ensemble size increases, instances with positive (negative) margin tend to be correctly (incorrectly) classified. Therefore, the asymptotic prediction error of the ensemble can be computed as the fraction of instances with positive margin.

The main difference with previous investigations is that in the present analysis we compare the prediction of a finite ensemble with the prediction of an ensemble of infinite size and not with the actual class of the instance to be classified. Targeting the infinite ensemble prediction has the advantage that it can be accurately estimated without knowledge of the true class labels. Therefore, it is not necessary to have access to the probability that an ensemble classifier predicts the correct class label for a particular instance. The estimation of these probabilities from the training data can be unreliable because of biases in the learning process.

The related problem of estimating the prediction of a finite ensemble by querying only a subset of classifiers was considered in [10]. In that work, the main goal was to reduce the number of classifiers needed for prediction (ensemble pruning). The current investigation provides an extension of the analysis presented in [10] to ensembles of infinite size.

4. Experiments

The application of the probabilistic framework for inference on the asymptotic prediction of parallel ensembles is illustrated in a variety of classification problems from the UCI repository [18]. The ensembles used for the empirical validation of this analysis are bagging [1] and random forest [2]. In bagging, the individual classifiers are built by applying the same learning algorithm to

independent bootstrap samples of the training set [1]. Each bootstrap sample has the same size as the original training set and is obtained by drawing with replacement from this set. Random forest [2] was introduced as an improvement over bagging when the classifiers of the ensemble are decision trees. Besides resampling, random forest uses randomized decision trees. The splits in the internal nodes of a randomized decision tree are made in terms of a subset of randomly selected attributes.

The characteristics of the datasets used in the analysis are displayed in Table 1. For each classification problem, the data are randomly partitioned into two disjoint sets. The first set, which contains two thirds of the available data, is used for training. The remaining data are used for testing. In the synthetic problems *Twonorm*, *Ringnorm*, *Threenorm*, *Led* and *Waveform* the training set (300 instances) and the test set (1000 instances) are generated by random sampling from the model distribution. This process is repeated 100 times for each dataset. The results reported are averages over these realizations.

Table 1: Datasets used in the experiments.

Problem	**Attributes**	**Instances**	**Classes**
breast	9	669	2
glass	9	214	6
heart	13	270	2
led	7	-	10
liver	6	345	2
new-thyroid	5	215	3
pima	8	768	2
ringnorm	20	-	2
spam	57	4,601	2
threenorm	20	-	2
twonorm	20	-	2
vehicle	18	846	4
vowel	10	990	11
waveform	21	-	3
wine	13	178	3

For each realization, a random forest (RF) [2] and a bagging ensemble

[1] composed of unpruned CART trees [19] are built using the corresponding training set. Both RF and bagging are parallel ensemble learning algorithms in which the individual classifiers are built independently when conditioned to the training data. Therefore, the framework introduced in Section 2 is appropriate to describe their prediction by majority voting.

The analysis focuses on test instances for which the querying process is halted when at most 101 trees have been polled. This particular value (101 classifiers) has been selected because it is a common choice for the ensemble size in the literature on bagging and random forests [2, 4, 20, 7]. An odd number of classifiers is used to avoid ties in binary classification problems. For these instances, the prediction of the finite ensemble is compared with the prediction of an ensemble of the same type, built using the same training set, and composed of $10,000$ trees. This ensemble is sufficiently large to serve as a proxy for the ensemble of infinite size because it assigns the asymptotic class label to all but a very small fraction of the test instances.

The evaluation protocol involves the following steps: For each instance in the test set, the classifiers in the ensemble are sequentially queried. The vector $\mathbf{t}$, which keeps a tally of the class predictions, is updated after the prediction a new classifier becomes known. This vector of votes is then used to compute the value of $\mathcal{L}(y^\star|\mathbf{t})$ for the provisional majority class. The voting process is stopped if $\mathcal{L}(y^\star|\mathbf{t}) > \pi$ with $\pi = 99\%$. For these instances, the class predicted by the partial ensemble is compared with the prediction of the ensemble composed of $10,000$ trees. Then, both predictions are compared with the true class label of the instance to estimate the corresponding generalization errors. We also record the percentage of instances whose asymptotic class label can be estimated at the specified level of confidence by querying at most 101 trees and the average number of classifiers involved in the prediction. Finally, we also determine whether using the same finite ensemble (irrespective of the instance considered), instead of selecting a different number of classifiers depending on the instance considered (instance-based strategy), also yields correct estimates of the asymptotic predictions with high probability. For this purpose, the results for an ensemble of fixed size are reported as well. The number of classifiers queried in this fixed-size ensemble (FS) strategy is the ceiling of the average number of classifiers obtained by the instance-based (IB) strategy. The computational cost of the FS strategy is similar to the cost of the IB strategy because the overhead to determine whether the voting process should be stopped is negligible compared to the time needed to query a classifier.

Tables 2 and 3 display the results of experiments on different benchmark classification problems for bagging and RF, respectively. The first column gives the percentage of test instances that receive a stable classification with a confidence level $\pi = 99\%$ by querying at most 101 classifiers. These percentages are fairly high for both RF and bagging. They range between $\approx 75\%$ and $\approx 98\%$ in the problems investigated. The remaining columns of these tables report averages over these test instances only. The third column presents the average number of classifiers queried in the voting process when the instance-based stopping criterion is applied. The average number of trees queried in the IB strategy varies between ≈ 8 and ≈ 25. Similar sizes are obtained using bagging and RF. These values are well below the maximum number of trees considered, which is 101. The fourth and the fifth columns display the average disagreement rates between the predictions of the finite ensembles (labeled Bag-FS and RF-FS for the strategy that uses a fixed number of classifiers, and Bag-IB and RF-IB for the strategy that determines the number of classifiers needed dynamically, depending on the instance considered) and the predictions of the corresponding asymptotic ensembles. The disagreement rates in the IB strategies are generally below the level of $1 - \pi = 1\%$ fixed in the experiments. This behavior is the result of the fact that $\mathcal{L}(y_i|\mathbf{t})$ is a lower bound and that the uniform prior for $\boldsymbol{p}$ in Eq.˜(5) generally produces conservative estimates for the stopping point. Slightly higher rates than 1% are observed in the classification problems *Liver* and *Threenorm*. In these problems, the distribution of $\boldsymbol{p}$ in the test data has a mode around $p_1 = 1/2$, which implies that the assumption of a uniform prior in (5) introduces a non-conservative bias in the estimation of the posterior distribution. The disagreement rates for the FS ensembles are much larger than for the corresponding IB ensembles. Furthermore, they are above the level of 1% in many of the problems investigated. This means that using a fixed ensemble size for all instances is not an effective strategy. The sixth and seventh columns display the average error rate of the different strategies. The last columns of each table, labeled RF∞ and Bag∞, respectively, present the error rates of ensembles of $10,000$ trees. Note that the differences between the error rates of the finite ensemble (columns 5 and 6) and the infinite ensemble (column 7) are necessarily smaller than the disagreement rates reported in columns 3 and 4. From these results it is apparent that the error differences between the IB ensembles and the asymptotic ones are almost negligible in most of the problems investigated. By contrast, the error rates of the finite FS ensembles are systematically worse than the corresponding infinite ensembles.

Table 2: Experimental results for bagging. Average fraction of test instances for which the prediction by the finite ensemble provides an estimate of the asymptotic prediction whose confidence level is above $\pi = 99\%$, average fraction of trees needed for these instances, average disagreement rates between the predictions of the finite ensembles and of the asymptotic one, and, finally, classification error of the different ensembles for these instances. The results are given for the instance-based (IB) and for the fixed-size (FS) strategies.

Problem	**% of test instances**	**# Trees Bag-IB**	**% of disagreement**		**Classification Error in %**		
			Bag-FS	**Bag-IB**	**Bag-FS**	**Bag-IB**	**Bag∞**
breast	97.8±0.9	8.1±0.6	0.8±0.6	0.1±0.2	3.6±1.1	3.2±1.0	3.2±1.0
glass	82.3±4.4	21.1±2.8	1.7±1.7	0.4±0.7	21.7±5.6	21.4±5.5	21.4±5.5
heart	87.4±4.0	16.7±1.9	1.4±1.4	0.4±0.8	16.5±4.5	16.2±4.2	16.2±4.3
led	91.4±2.6	13.8±1.4	1.0±1.0	0.1±0.3	26.0±2.1	25.8±2.1	25.8±2.1
liver	78.8±4.2	23.9±2.5	2.3±1.5	1.1±1.2	25.6±3.5	25.4±3.5	25.4±3.7
new-thyroid	97.0±2.1	9.7±1.5	0.8±1.2	0.1±0.4	4.8±2.6	4.7±2.6	4.7±2.7
pima	85.0±2.6	18.1±1.4	2.2±1.0	0.7±0.6	21.4±2.5	20.8±2.6	20.8±2.6
ringnorm	87.2±1.8	19.6±1.6	1.8±0.6	0.5±0.3	6.4±1.7	5.5±1.7	5.4±1.7
spam	96.9±0.5	9.5±0.3	1.0±0.3	0.1±0.1	5.2±0.6	4.8±0.6	4.8±0.6
threenorm	76.5±1.9	25.5±1.5	2.3±0.8	1.1±0.5	13.8±2.2	13.1±2.2	12.9±2.2
twonorm	88.8±1.1	18.4±0.7	1.7±0.5	0.4±0.3	4.4±1.2	3.4±1.2	3.3±1.1
vehicle	80.2±2.7	20.7±1.5	1.8±0.9	0.5±0.5	18.7±2.5	18.4±2.5	18.4±2.5
vowel	85.7±1.9	23.2±1.1	1.1±0.6	0.1±0.2	5.8±1.6	5.1±1.5	5.0±1.5
waveform	83.4±1.8	20.6±1.3	1.9±0.5	0.5±0.3	15.1±1.9	14.6±1.9	14.6±1.8
wine	96.4±2.7	11.3±1.7	1.1±1.4	0.1±0.5	3.4±2.7	2.8±2.5	2.8±2.5

Table 3: Experimental results for random forest (RF). Average fraction of test instances for which the prediction by the finite ensemble provides an estimate of the asymptotic prediction whose confidence level is above $\pi = 99\%$, average fraction of trees needed for these instances, average disagreement rates between the predictions of the finite ensembles and the asymptotic one, and finally, classification error of the different ensembles for these instances. The results are given for the instance-based (IB) and the fixed-size (FS) strategies.

Problem	**% of test instances**	**# Trees RF-IB**	**% of disagreement RF-FS**	**RF-IB**	**Classification Error in % RF-FS**	**RF-IB**	**RF∞**
breast	98.0±0.8	8.1±0.5	0.7±0.5	0.1±0.2	3.1±1.0	2.8±0.9	2.7±0.9
glass	78.9±4.8	22.9±2.6	1.2±1.5	0.3±0.6	18.1±5.5	17.6±5.4	17.6±5.4
heart	86.1±3.1	18.6±2.5	2.1±1.6	0.6±0.9	14.4±3.6	13.7±3.6	13.6±3.5
led	82.7±4.9	23.9±2.9	1.0±1.3	0.2±0.6	22.5±2.1	22.2±2.0	22.2±2.0
liver	75.1±4.2	26.7±2.7	2.6±1.8	1.2±1.4	24.5±4.2	23.8±4.2	23.4±4.1
new-thyroid	96.1±2.5	10.6±1.2	1.0±1.3	0.1±0.3	3.3±2.2	2.9±2.1	2.9±2.1
pima	83.6±2.3	20.0±1.5	2.1±1.0	0.7±0.5	20.6±2.2	20.3±2.4	20.2±2.4
ringnorm	88.3±1.3	20.1±1.3	1.8±0.5	0.5±0.3	4.3±1.0	3.3±0.9	3.2±0.9
spam	96.5±0.4	10.3±0.3	1.0±0.3	0.1±0.1	4.2±0.5	3.7±0.5	3.7±0.5
threenorm	73.8±1.8	27.6±1.2	2.4±0.6	1.3±0.5	11.4±1.5	10.6±1.4	10.3±1.4
twonorm	90.2±1.0	18.7±0.7	1.5±0.4	0.4±0.2	2.9±0.5	1.9±0.5	1.7±0.5
vehicle	77.5±2.7	22.0±1.3	1.8±0.9	0.5±0.5	16.4±2.6	16.2±2.6	16.2±2.6
vowel	86.2±2.1	25.8±1.1	0.9±0.6	0.1±0.2	2.7±1.0	2.0±1.0	2.0±1.0
waveform	80.5±1.7	23.8±1.1	1.8±0.5	0.7±0.3	11.9±1.3	11.4±1.2	11.3±1.3
wine	97.1±2.0	12.2±1.5	0.9±1.3	0.1±0.3	1.9±1.7	1.3±1.4	1.3±1.4

The performance of the asymptotic and the finite ensembles are compared using the statistical framework introduced in [21]. This framework allows to compare the overall performance of the different classification systems in several problems. In general, comparisons across multiple datasets are less affected by Type 1 errors than comparisons that use a single dataset. This is because in the former the variance comes from the differences between the data sets, which are generally independent. By contrast, when a single dataset is used for comparison, the variance of the results comes from variations among different partitions of the same data. Because of the dependencies among the different partitions, this variance is typically underestimated. To perform the comparison, the different methods are ranked according to their performance in each of the problems considered (rank 1 for the best method, rank 2 for the second best and so on). Then, the average of the ranks obtained by each method in each of the problems is computed. Finally, statistical tests are applied to determine whether the differences among the average ranks of the methods considered are statistically significant. Robust, non-parametric tests are used because many of the assumptions made by standard parametric tests are often violated when analyzing the performance of machine learning algorithms [21]. In these tests, RF and bagging ensembles are analyzed separately. A Friedman test based on these average ranks rejects (with a p-value < 0.05) the null-hypothesis that there are no significant differences in performance among the different methods evaluated, for both bagging and RF. Finally, a Nemenyi test is applied to determine whether the differences in average rank are statistically significant. If the average ranks of the methods differ by more than a critical distance (CD), the differences are statistically significant.

Fig. 5 displays the results of this test for both bagging (top) and RF (bottom). Ensembles for which the differences in average rank are not statistically significant with a p-value < 0.05 are connected with a horizontal solid black line. The critical distance (CD) above which the differences in average rank are considered significant is displayed at the top of each plot. From the results of this test one can conclude that, for both RF and bagging ensembles, there are no significant differences in performance between the IB and the asymptotic ensembles when only the instances that can be assigned a final class label with at most 101 classifiers are considered. By contrast, the deterioration in performance caused by using an ensemble of fixed size is statistically significant.

Finally, we carry out additional experiments to investigate the depen-

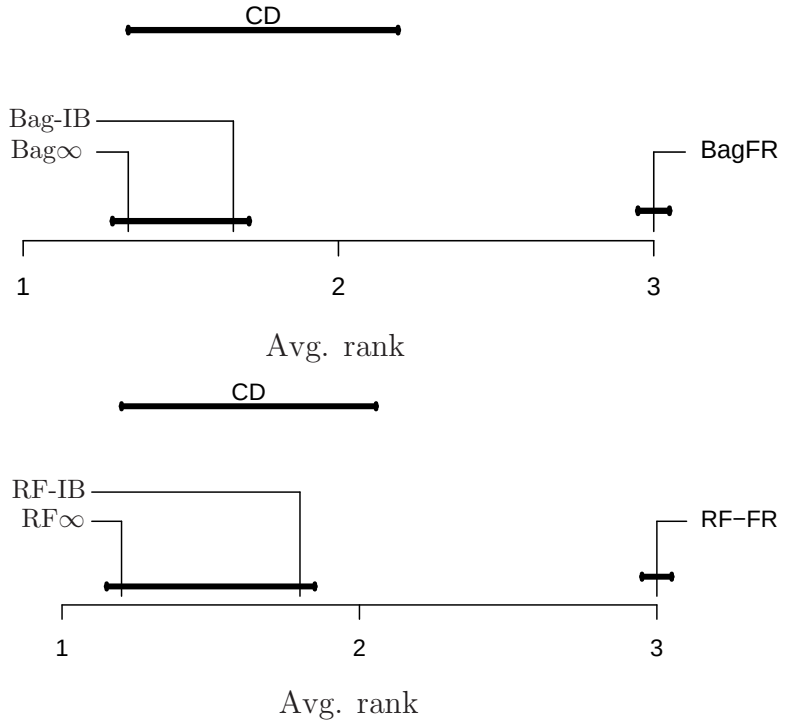

Fig. 5: Results of a Nemenyi test on the average ranks of the finite and the asymptotic ensembles in the classification problems investigated for bagging (top) and RF (bottom). Ensembles for which the differences in average rank are not statistically significant with p-value < 0.05 are linked with a solid black line. The critical distance (CD) above which the differences in rank are considered significant is displayed at the top of each figure.

dence on π of the fraction of test instances that satisfy the stopping criterion before querying the 101 classifiers in the ensemble. The curves that display the dependence of the fraction of such examples as a function of π are plotted in Fig. 6 for the classification problems *Pima* and *Waveform*. These curves are representative of all the classification problems investigated. They show that there is a trade-off between the desired level of confidence in the predictions (π) and the number of test instances for which it is sufficient to query at most 101 classifiers to obtain a stable prediction at the specified confidence level. The larger the value of π, the lower the number of instances that reach this confidence level. Specifically, when π approaches 100% the fraction of these instances tends to zero. As the confidence level is decreased, but still remains close to 100%, there is a sharp increase of the fraction of test in-

stances whose classification with 101 trees is stable. Eventually, the curves saturate and only small benefits are obtained by allowing a lower confidence level on the estimates.

5. Conclusions

In this paper we have introduced a probabilistic framework for making inference on the asymptotic (infinite) ensemble prediction. For this, we have used the evidence given by the output of a finite set of ensemble classifiers. The analysis presented is based on the estimation of the class prediction probabilities of a single classifier for a given test instance. To estimate these probabilities, the classifiers in the ensemble are queried sequentially. Starting from an uniform prior, Bayes' theorem is used to update the probability estimates as new predictions become known. These estimates are used to compute the probability that the provisional majority class (determined on the basis of the known class votes) coincides with the asymptotic ensemble prediction. Since the evaluation of this probability is costly for multi-class classification problems, we use an approximation based on a lower bound that can be readily computed.

The framework considered can be used to identify data instances for which the predictions of a finite ensemble are sufficient to estimate the asymptotic class prediction with a high level of confidence. The analysis applies to collections of classifiers trained on independent realizations of a randomized learning algorithm and whose predictions are combined by majority voting. To classify a test instance, the individual classifiers from the ensemble are queried sequentially. After evaluating the output of each classifier, we compute an estimate of the probability that the provisional majority class coincides with the asymptotic ensemble prediction. The estimate (actually, a lower bound of this probability) is computed in terms of the predictions of the classifiers that have been queried up to that moment. If the lower bound is above a specified threshold π, the voting process can be stopped. Given that the actual probability is approximated using a lower bound, the voting process does not stop prematurely in most cases. For instances in which the querying process is halted, the predictions of the partially queried ensemble and the predictions of an ensemble of infinite size are expected to coincide with a probability larger than $\approx \pi$. Furthermore, the disagreement rate between the finite ensemble and the asymptotic one is bounded from above by $1 - \pi$.

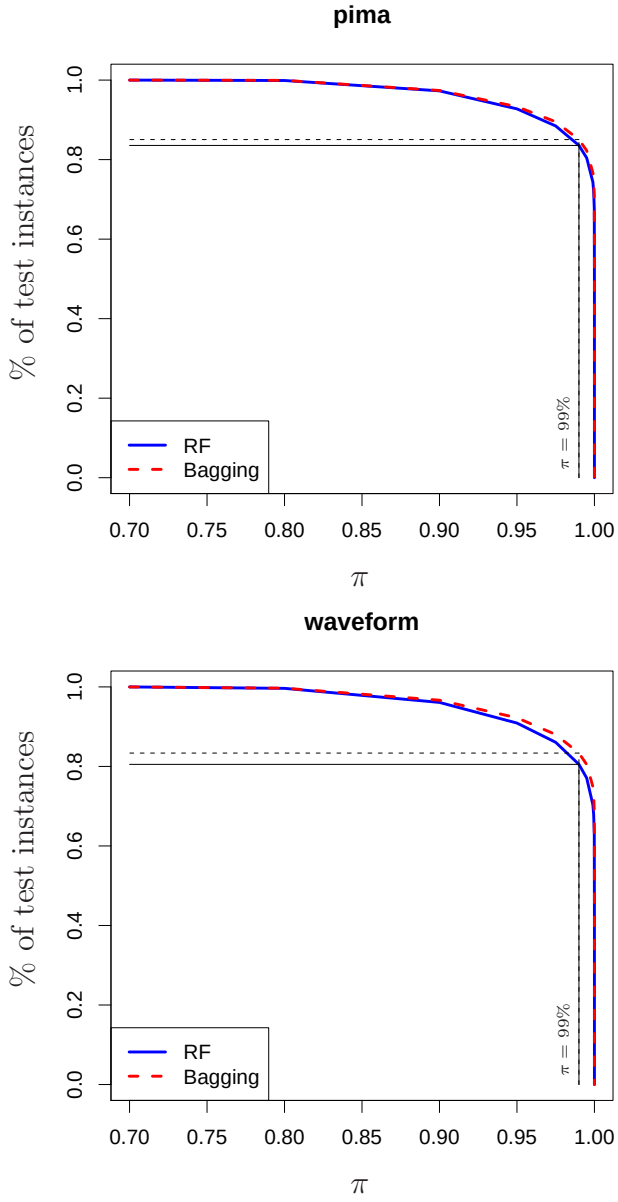

Fig. 6: Fraction of test instances in which ensembles of at most 101 classifiers reach a confidence level in their estimate of the asymptotic error that is above π for different values of π. Values corresponding to $\pi = 99\%$ are marked with a dotted line in the case of bagging, and with a solid line in the case of RF.

An experimental evaluation in a variety of binary and multi-class classification problems illustrates the application of this framework to describe majority voting in random forest and bagging ensembles. In the problems investigated, a large fraction of the test instances require on average a fairly small number of classifiers to gather sufficient statistical evidence on the asymptotic ensemble prediction. For these instances, the disagreement rates between the predictions of the finite ensemble and the asymptotic ones are close to, and in most cases below the specified confidence level. Furthermore, the differences between the generalization error of the finite ensembles on these instances and the asymptotic infinite-ensemble limit are much smaller than the differences in classification. This means that the changes in the class labels that arise if the querying process is continued, affect both correctly and incorrectly classified instances in approximately equal numbers.

Acknowledgements

The authors acknowledge support from the Spanish Ministerio de Ciencia e Innovación, projects TIN2007-66862-C02-02 and TIN2010-21575-C02-02.

References

[1] L.˜Breiman, Bagging predictors, Machine Learning 24˜(2) (1996) 123–140.

[2] L.˜Breiman, Random forests, Machine Learning 45˜(1) (2001) 5–32.

[3] T.˜G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, Machine Learning 40˜(2) (2000) 139–157.

[4] P.˜Buhlmann, Bagging, subagging and bragging for improving some prediction algorithms, in: M.˜Akritas, D.˜Politis (Eds.), Recent Advances and Trends in Nonparametric Statistics, 2003, pp. 19–34.

[5] G.˜Martinez-Muñoz, A.˜Suárez, Switching class labels to generate classification ensembles, Pattern Recognition 38˜(10) (2005) 1483–1494.

[6] J.˜J. Rodríguez, L.˜I. Kuncheva, C.˜J. Alonso, Rotation forest: A new classifier ensemble method, IEEE Transactions on Pattern Analysis and Machine Intelligence 28˜(10) (2006) 1619–1630.

[7] P.˜Geurts, D.˜Ernst, L.˜Wehenkel, Extremely randomized trees, Machine Learning 36˜(1) (2006) 3–42.

[8] D.˜Opitz, R.˜Maclin, Popular ensemble methods: An empirical study, Journal of Artificial Intelligence Research 11 (1999) 169–198.

[9] L.˜Hansen, P.˜Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis and Machine Intelligence 12˜(10) (1990) 993–1001.

[10] D.˜Hernández-Lobato, G.˜Martínez-Muñoz, A.˜Suárez, Statistical instance-based pruning in ensembles of independent classifiers, IEEE Transactions on Pattern Analysis Machine Intelligence 31˜(2) (2009) 364–369.

[11] M.˜Abramowitz, I.˜A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, Dover, New York, 1964.

[12] C.˜M. Fortuin, P.˜W. Kasteleyn, J.˜Glnibre, Correlation inequalities on some partially ordered sets, Comm. Math. Phys 22 (1971) 89–103.

[13] Y.˜Freund, R.˜E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: Proc. 2nd European Conference on Computational Learning Theory, 1995, pp. 23–37.

[14] Y.˜Freund, R.˜E. Schapire, Experiments with a new boosting algorithm, in: International Conference on Machine Learning, 1996, pp. 148–156.

[15] R.˜Esposito, L.˜Saitta, Experimental comparison between bagging and Monte Carlo ensemble classification, in: ICML '05: Proceedings of the 22nd international conference on Machine learning, ACM Press, New York, NY, USA, 2005, pp. 209–216.

[16] R.˜Esposito, L.˜Saitta, A Monte Carlo analysis of ensemble classification, in: R.˜Greiner, D.˜Schuurmans (Eds.), Proceedings of the twenty-first International Conference on Machine Learning, ACM Press, New York, NY, Banff, Canada, 2004, pp. 265–272.

[17] Q.˜Cai, C.˜Zhang, C.˜Peng, Analysis of classification margin for classification accuracy with applications, Neurocomputing 72˜(7-9) (2009) 1960 – 1968.

[18] A.˜Asuncion, D.˜Newman, UCI machine learning repository, http://www.ics.uci.edu/˜mlearn/MLRepository.html (2007).

[19] L.˜Breiman, J.˜H. Friedman, R.˜A. Olshen, C.˜J. Stone, Classification and Regression Trees, Chapman & Hall, New York, 1984.

[20] G.˜Martínez-Muñoz, D.˜Hernández-Lobato, A.˜Suárez, An analysis of ensemble pruning techniques based on ordered aggregation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 245–259.

[21] J.˜Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.